CSE 151 Intro. to AI: A Statistical Approach

Instructor: Kamalika Chaudhuri
CSE 151 Machine Learning

Instructor: Kamalika Chaudhuri
Course Staff

**Instructor:** Kamalika Chaudhuri (kamalika@cs)

**Office Hours:** MW, 2-3pm, 4110

**TAs:**

Yuncong Chen (yuc007@eng.ucsd.edu)

**Office Hours:** F 3-4pm, B250A

Qiushi Wang (qiushi@ucsd.edu)

**Office Hours:** T 10-11am, B250A

Sahil Goyal (sagoyal@cs.ucsd.edu)

**Office Hours:** Th 3:30-4:30pm, B250A

Wangfan Fu (wafu@eng.ucsd.edu)
What is Machine Learning?

How to use data to learn to make better predictions

Example 1: Recommendation Systems
What is Machine Learning?

How to use data **to learn** to make better predictions

Example 2: Spam Detection
What is Machine Learning?

How to use data **to learn** to make better predictions

Example 3: Link Prediction
What is Machine Learning?

How to use data to learn to make better predictions
Algorithm behavior changes based on data

This class: some basic machine learning methods
Two Types of Machine Learning

**Supervised Learning**
Given examples of data and their labels, predict labels of new (unseen) data

**Unsupervised Learning**
Given data, build a model or cluster

There are other types, but we won’t get to it in this class
Supervised Learning

Classification:

Given labeled data:

\[ (x_i, y_i) \quad i=1,\ldots,n \]

where \( y \) is \textit{discrete}, find a rule to predict \( y \) values for \textit{unseen} \( x \)
Typical Classification Algorithm

Set of input examples \((x_i, y_i)\)

Classification Algorithm

Prediction Rule

New example \(x\)

Label \(y\)
Typical Classification Algorithm

Set of input examples \((x_i, y_i)\)

Classification Algorithm

Prediction Rule

New example \(x\)

Label \(y\)

Training and test data must be separate!
Typical Classification Algorithm

Set of input examples \((x_i, y_i)\)

Classification Algorithm

Prediction Rule

Test Data

New example \(x\)

Label \(y\)

Performance Measure:
Accuracy (or fraction of correct answers) on test data
Supervised Learning

Classification: Given labeled data \((x_i, y_i)\) where \(y\) is **discrete**, predict \(y\) values for unseen \(x\)

**Example 1:** Predict if a **new** patient has flu or not, based on **existing** patient data

What is \(x\) and \(y\)?
Supervised Learning

**Classification:** Given labeled data \((x_i, y_i)\) where \(y\) is discrete, predict \(y\) values for unseen \(x\)

**Example 1:** Predict if a patient has flu or not

<table>
<thead>
<tr>
<th>Fever</th>
<th>Cold</th>
<th>Temperature</th>
<th>Flu?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
<td>99F</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Features: Properties of patient
Label: Flu/No flu

A **binary** (two-label) classification problem
**Supervised Learning**

**Classification:** Given labeled data \((x_i, y_i)\)
where \(y\) is **discrete**, predict \(y\) values for unseen \(x\)

Example 2: Which digit in the image?

![Digits Image]

Label: 0, 1, ..., 9

What are the features?

A **multiclass** classification problem
Supervised Learning

**Classification:** Given labeled data \((x_i, y_i)\)
where \(y\) is **discrete**, predict \(y\) values for unseen \(x\)

**Example 2:** Which digit in the image?

![Digits Image]

Label: 0, 1, ..., 9

What are the features?
Option: vector of pixel colors

- Image
- \(x\) (0 for white, 1 for black)
Supervised Learning

**Classification:** Given labeled data \((x_i, y_i)\)
where \(y\) is **discrete**, predict \(y\) values for unseen \(x\)

**Example 2:** Which digit in the image?

```
0 1 2 3 4
5 6 7 8 9
```

Label: 0,1,...,9

What are the features?
Option: vector of pixel colors

There are other options too

**Lesson:** Choosing features is non-trivial in real applications
Supervised Learning

**Classification:** Given labeled data \((x_i, y_i)\)
where \(y\) is **discrete**, predict \(y\) values for unseen \(x\)

**Example 3:** Spam or not?

**Email 1**
From: Canadian Pharmacy 
Subject: Offer ends now!

**Email 2**
From: Yuncong Chen 
Subject: TA meeting

<table>
<thead>
<tr>
<th>Pharmacy</th>
<th>offer</th>
<th>meeting</th>
<th>TA</th>
<th>Spam?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email 1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Email 2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Label: 0 (not spam), 1 (spam) 
Features: Words in the email
Supervised Learning

Regression:

Given data:

\[(x_i, \ y_i) \quad i = 1, \ldots, n\]

where \(y\) is continuous, design a rule to predict \(y\) values for unseen \(x\)
Supervised Learning

**Regression:** Given data \((x_i, y_i)\)
where \(y\) is **continuous**, predict \(y\) values for unseen \(x\)

**Example 1:** Predict house price from properties of house

<table>
<thead>
<tr>
<th>Bedrooms</th>
<th>Bathrooms</th>
<th>Area</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>2000</td>
<td>600K</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1200</td>
<td>400K</td>
</tr>
</tbody>
</table>

Independent Variable: Property of house
Dependent variable: price
Two Types of Machine Learning

**Supervised Learning**
Given examples of data and their labels, predict labels of new (unseen) data
Examples: Classification, Regression

**Unsupervised Learning**
Given data, build a model

There are other types, but we won’t get to it in this class
Unsupervised Learning

Clustering
Given a set of input objects, group them to clusters by similarity

Example 1: Cluster videos by people in them
Unsupervised Learning

Clustering
Given a set of input objects, group them to clusters by similarity

Example 2: Cluster documents by topic

Physics
- Gravity
- Laws of Motion
- Electricity

Math
- Geometry
- Algebra

Features: Words in the document
Unsupervised Learning

Dimensionality Reduction
Given high dimensional data, find a good low dimensional representation

Example 1: Images

0 1 2 3 4
5 6 7 8 9

Number of pixels = 768, so 768-dimensional object
Can we find a lower dimensional representation?
Two Types of Machine Learning

**Supervised Learning**
Given examples of data and their labels, predict labels of new (unseen) data
Examples: Classification, Regression

**Unsupervised Learning**
Given data, build a model
Examples: Clustering, Dimension Reduction, learning HMMs

There are other types, but we won’t get to it in this class
Logistics

Instructor: Kamalika Chaudhuri
Email: kamalika@cs.ucsd.edu

Lecture: TTh 5-6:20pm, WLH 2005
Sections (optional): W8-8:50am PETER 103
W2-2:50pm WLH 2205

Website: http://cseweb.ucsd.edu/classes/sp13/cse151-a/
Logistics

The class is currently at capacity

I will not sign any add-cards
Administrivia

Textbooks:
No textbook for this class

Syllabus:
Classification -- k-NN, Perceptron, Boosting, etc.
Linear Least Squares Regression
Unsupervised learning -- k-means, hierarchical clustering
Prerequisites

**Probability:** Events, random variables, expectations, joint, conditional and marginal distributions, independence

**Linear Algebra:** Vector spaces, subspace, matrix inversion, matrix multiplication, linear independence, rank, determinant, bases, orthonormality, solving systems of linear equations

**Calculus:** Minima, maxima of functions, derivatives, integrals

**Programming:** Write programs in a language of your choice. No hand-holding provided
Prerequisites

Calibration homework HW0 is out!

Due in lecture on April 11

HW0 covers most (but not all) of the material you need to know as a pre-requisite
Assessment

Homeworks (7): 30%
Midterm: 30%
Final: 35%
Class Participation: 5%

We will be using GradeSource
Homework Policy

Homeworks are due *in class* at the beginning of lecture

No late homeworks will be accepted

Homework with the lowest grade will be dropped

Homeworks will be graded based on *correctness* and *clarity*
Homeworks

Lectures will cover background conceptual material you need to do your assignments

Homeworks will be a mix of programming + pen and paper assignments

You can use any language and any libraries for your programming assignments

If you use external libraries, it is your responsibility to make sure they give you correct answers

Submit a printout of your code with your homework
Collaboration Policy

Homeworks should be done in **groups of one or two**

You are not permitted to collaborate with anyone outside your homework group

Email me the name of your homework partner by April 11

If you need a homework partner, please post on Piazza
Exam Policy

Midterm and final will be closed book, closed notes

You can bring a “crib-sheet” with you
Regrade Policy

We will consider regrade requests for 7 days from when the graded homeworks/midterm is given out.

For a **midterm regrade request**, I will need a one-page written report. A regrade request will not be considered without a report.

More details on the class web-site.
Class Participation

Class participation is very important for this class

In some lectures, we will have mini problem-sessions

Your class participation grade depends on how you do in those sessions

Bring blank sheets of paper to class for problem sessions

Ask lots of questions, if you don’t understand something!
Questions

Message board for this class on Piazza

Please post your questions on the message board!
Academic Honesty

**Remember:**
Academic dishonesty is taken very seriously in this class

More details on the class website
Typical Classification Algorithm

Training Data

Set of input examples \((x_i, y_i)\)

Classification Algorithm

Prediction Rule

Test Data

New example x

Label y

Training and test data must be separate!
Generative Classification

Goal: Classify red from blue

Generative:
Model each class probabilistically
Learn the parameters of each class from data
Generative Classification

Goal: Classify red from blue

Generative:
Model each class probabilistically
Learn the parameters of each class from data

For a test example x, find $P(\text{class 1}|x)$ and $P(\text{class 2}|x)$
Report 1 if $P(\text{class 1}|x) > P(\text{class 2}|x)$, 2 otherwise
Discriminative Classification

**Goal:** Classify red from blue

**Discriminative:**
- No need to model each class probabilistically
- Find a suitable separator (say a linear separator) that mostly separates red from blue

Advantages and Disadvantages?
Generative vs. Discriminative

This class we will mostly cover discriminative models

A few generative models at the end if time permits